DRAGON SYSTEMS' 1997 BROADCAST NEWS TRANSCRIPTION SYSTEM

Steven Wegmann, Francesco Scattone,
Ira Carp, Larry Gillick, Robert Roth, Jon Yamron

Dragon Systems, Inc.
320 Nevada Street, Newton, MA 02160

1. INTRODUCTION

This system represents Dragon's first participation in the HUB4 evaluations since the 1995 Marketplace dry run. At that time, we used a fairly complicated system which had three sets of acoustic models: one for clean wide-bandwidth data, one for low-bandwidth data, and one for speech with music in the background. Our system produced small pieces that were labelled by channel type and then decoded with the appropriate model set [1]. Our 1997 evaluation system is much simpler, since we use one set of gender independent, speaker normalized models to recognize all of the data. In the two years between the dry run and the current evaluation, much of our development work focused on the Switchboard corpus [2], including many techniques -- such as speaker normalization and rapid adaptation -- which now make it possible to consolidate the treatment across channels and speakers. The current evaluation system in many ways represents the transfer of these new techniques into our Broadcast News recognizer together with building the infrastructure necessary to handle the HUB4 data.

In the sections that follow, we describe the HUB4 evaluation system and then discuss several families of experiments exploring the performance of key components of the system.

2. SYSTEM OVERVIEW

Dragon's continuous speech recognizer is a time synchronous Hidden Markov Model based sytem, which has been described extensively elsewhere ([1], [3]).

A total of 36 parameters are computed every 10 milliseconds: 12 cepstral parameters, 12 cepstral differences, 12 cepstral second differences. We use PLP-based cepstra [4], computed in the style of Cambridge/HTK, as reported in [5]. This set of 36 parameters is linearly transformed using IMELDA techniques [6] to a set of 24 parameters which are used for training and recognition. Speaker normalization ([3], [7]) is used to reduce variability among speakers due to vocal tract length. During the signal processing stage, the frequency scale is "warped" using a piecewise linear transformation.

The model for a sentence hypothesis is obtained by concatenating models we call PICs (for "phonemes-in-context"). For this evaluation, we used triphone models trained from the HUB4 acoustic training data plus data drawn from the Wall Street Journal, Marketplace, and WSJCAM0 corpora. A 51-element phoneme set was used that has syllabic consonants and two stress levels for certain vowels. PICs are modeled using from 3 to 4 nodes, with each node having an output distribution (PEL) and a duration distribution. Which PEL model to employ for any given position of any PIC is determined based on a decision tree whose nodes ask linguistic questions about neighboring phonemes as well as questions about the position of word boundaries. The PEL models themselves are general mixture models with basis components given by multivariate gaussian distributions with diagonal covariance.

In addition to speaker normalization, we also make use of rapid adaptation, using linear regression techniques to construct transformations of acoustic parameter space mapping speaker-independent model means to speaker-specific ones. This approach was inspired by, and represents a simplification of, speaker adaptation strategies implemented by Cambridge [8]. We also used speaker adaptation techniques (SAT) during training ([2], [9]): Training speech is force-aligned to transcripts and the usual adaptation transformations are computed mapping speaker-independent model to speaker-specific data, and then a sort of "inverse" transformation is performed on the speech frames. This permits the training of new models with the transformed data which should behave well under test-time adaptation. We used 4 transformation classes both at training and test, determined by grouping related phonemes. (For another approach to speaker-adaptive training, see [10].)

The evaluation system used a three-way interpolated trigram language model and 57K vocabulary, described in more detail in section 3.1 below.

To deal with the huge unbroken audio stream, we developed a system to transform the input into clusters of smaller more digestible pieces. Since the clusters roughly correspond to speakers, we treat them as such when we choose frequency warps for speaker normalization and when we perform rapid adaptation.

Here are the basic steps in the system protocol:

- An amplitude-based silence detector is used to break the input into chunks that are 20 to 30 seconds long.
- A phoneme recognizer is used to produce a more refined chopping of these chunks.
- The segments are clustered for speaker normalization and unsupervised adaptation.
- Channel normalization is performed on each segment.
- Speaker normalization is performed within each cluster by doing a quick, errorful recognition with small acoustic models (5000 PELs) and a small bigram language model (300,000 bigrams), and then rescoring this transcript against speech data processed at each warp scale in order to pick the best scoring scale.
- Speaker normalized, SAT models with 12,000 PELs are used along with an interpolated trigram language model to obtain an initial transcription for each cluster.
- Unsupervised rapid adaptation is performed within each cluster, using 4 transformations, followed by the final recognition pass using the adapted acoustic models and the same trigram language model.

Because these are new aspects of the system, we provide somewhat more detail on the segmentation and clustering stages:

Phoneme Recognizer based Segmentation

We used a phoneme recognizer to segment the data into segments of reasonable size for recognition. The phoneme recognizer uses small gender independent acoustic models (5000 PELs) trained form the 1996 HUB4 acoustic training corpus and a tri-phone language model trained from Wall Street Journal data. We included pure music in the data used to train the silence model in the hope that silence would absorb segments containing pure music ([11]). A dynamic programming algorithm was then applied to chop the phone-labelled data: no segment was allowed to be shorter than 2 seconds or longer than 30 seconds, a penalty was applied if a break occurred in speech, and a boost was applied if a break occurred at the boundary of a region of silence with 4 or more consecutive frames.

Clustering

Here our goal is to construct clusters for which features such as channel conditions or the speakers' warp factors are comparable for all segments belonging to a given cluster. To do the clustering, we use the following measure of the distance from a segment s to a cluster c:

$$\mathrm{KL}(s, c+s) + \mathrm{KL}(c, c+s) + \mathrm{TimePen}(c, s)$$

where KL(a,b), the Kullback-Leibler distance, is the expectation under a of the logarithm of the ratio of the probability of the a distribution to the probability of the b distribution, and TimePen(a,b) is a linear function of the smallest time difference between a frame in a and a frame in b, truncated at a maximum value. The first term is a measure of how far a segment is to a cluster with the segment added. The second term is a measure of how far the cluster is to the same cluster with the segment added in. In some sense, this term measures how "unhappy" the other segments in the cluster would be if we add the segment in question, causing the whole cluster to move. The last term introduces a bias in favor of clustering together segments that are close in time.

Given the above distance measure, the clustering algorithm proceeds as follows. For each segment, we find the existing cluster with the minimum distance. If this distance is greater than a certain threshold (or if, as at the beginning, there are no existing clusters), we create a new cluster and remove the segment from the cluster it is in (if any). Otherwise, if the segment is not already in the closest cluster, we add it to that cluster and remove it from the cluster it is in. We iterate in this way through the segments a specified number of times. Optionally, the program does one final pass in which segments that belong to clusters with fewer than a specified number of frames are reassigned, thus eliminating clusters with too few frames.

We developed our clustering algorithm using gender independent, unwarped models. For these models, we produced clusters that worked as well as those created using the known speaker side information and the algorithm seemed relatively insensitive to threshold choices across a wide range. Explorations with more evaluation-like warped models are described in section 3.2 below.

3. EXPERIMENTS & ANALYSIS

3.1 Language Modelling

For the evaluation we used an interpolated language model. Three backoff trigram language models were trained from 500 million words of text.

- The Broadcast News acoustic training transcripts plus the 1995 Marketplace development transcripts were used to train a language model (A) with 300,000 bigrams and 640,000 trigrams.
- The Broadcast News language model training corpus was used to train a language model (B) with 7.1 million bigrams and 9.9 million trigrams.
- The 1995 Hub4 and Hub3 newswire texts were combined with 190 million words of commercially available newspaper data collected from the period January 1995 through June 1996, to train a language model (C) with 8.8 million bigrams and 14.9 million trigrams. We processed

data from the *Boston Globe*, the *Dallas Morning News*, the *Detroit Free Press*, the *Miami Herald*, *New York Newsday*, the *Philadelphia Inquirer*, and the *San Francisco Chronicle*.

All bigrams and trigrams were retained for set A, all bigrams but only trigrams occurring at least twice for B, and bigrams occurring at least twice and trigrams at least three times for C. These language models were interpolated with weights 0.22 for A, 0.50 for B, and 0.28 for C, obtained by minimizing perplexity on the 1996 HUB4 dev and eval tests.

The three language models share a 57K vocabulary constructed from the combined training sources. For the purpose of creating the vocabulary, the sources were combined with weights determined based on preliminary recognition runs on the 1996 devtest: 0.30 for A, 0.50 for B, and 0.20 for C. The top 57K words from this weighted collection were retained (62K including alternate pronunciations). This 57K set resulted in an OOV rate of 0.7% on the 1997 evaluation test.

How much did we gain from interpolating language model probabilities rather than simply merging all of the text sources, and how much did we gain from the addition of newspaper data? To attempt to answer the first question, we merged the counts from the Broadcast News language model training corpus, the newswire text, and the newspaper data. We then multiplied the counts from the acoustic training transcripts by a factor *m*, and merged them with the counts from the other corpora. From this text, we built a backoff trigram language model (retaining bigrams that occurred two or more times and trigrams that occurred three or more times), and computed the perplexity on the 1996 evaluation test set. The resulting perplexities are given in Table 1 along with the perplexity or the original interpolated language model.

m	*perplexity*
3	168.0
10	166.7
30	168.4
200	189.8
interp	148.3

Table 1: Language model perplexity for various weighted combinations of training data.

We re-ran the evaluation with the language model built from the data merged with $m = 10$. As you can see in Table 2, the interpolated (Eval) language model outperforms this "merged" model by a small amount. However, this study could hardly be called exhaustive. In particular we have not yet explored weighting the counts from the Broadcast News corpus, largely because of the expense involved in running these experiments in terms both of time and of diskspace. In this respect, experimenting with interpolated forms is much simpler, easily allowing us to sweep out interpolation weights without requiring the full rebuild of the language models being studied.

We also built a language model from merged counts, but now excluding the newspaper data used in source C. Again we see a small degradation when re-running the evaluation (the "-Papers" column).

	Eval	Merged	-Papers
F0	14.7	14.6	14.9
F1	24.4	25.3	25.8
F2	32.7	34.2	34.1
F3	35.0	37.0	37.6
F4	32.7	30.2	30.6
F5	20.1	21.1	22.3
FX	49.7	46.7	48.3
Total	23.4	24.0	24.4

Table 2: Performance of interpolated language model vs. models built from merged counts. (Figures give word error rate.)

3.2 Front End

We next turn to the contribution of frequency warping and PLP-based cepstra to our evaluation system. All of the models described in this section were trained from the 1996 HUB4 acoustic training corpus (~35 hours), and use the merged trigram language model described in the preceding section. Also, all of the acoustic models are about the same size, ~5000 PELs, smaller than our full evaluation models. The test data is the 1997 evaluation set using the same fixed chopping and clustering determined by the evaluation system. As in the evaluation, we performed rapid unsupervised adaptation with four transformations. All experiments use matched training and test, e.g. acoustic models trained from warped data are used when recognizing warped test data.

	PLP		MFCC	
	before	*after*	*before*	*after*
F0	18.7	17.4	18.6	17.7
F1	31.9	29.6	31.7	29.9
F2	48.1	41.4	47.2	42.1
F3	41.7	39.5	40.4	38.6
F4	37.2	33.3	36.3	34.2
F5	35.1	31.6	31.6	31.2
FX	55.8	54.4	55.9	54.0
Total	30.9	28.1	30.5	28.6

Table 3: Performance of MFCC vs. PLP cepstra both before and after adaptation.

First we compare our traditional melscale filterbank cepstral coefficients (MFCC) to the PLP cepstral coefficients using gender independent models (with no frequency warping). The resulting word error rates are reported in Table 3. Surprisingly, the traditional front-end does better prior to adaptation, but the PLP front-end does better after adaptation. We hypothesized that this might be because the clusters for adaptation were constructed using the PLP-processed data. However, when we tried using the traditional front-end when producing clusters for adaptation with traditional front-end based models in a follow-up experiment, we saw very little change in the results.

Next, we compared gender independent (GI-NW) and gender dependent (GD-NW) non-warped systems (using the PLP frontend) to a gender independent frequency warped system (GI-W) -- see Table 4. In the GD-NW system, gender detection was performed on a per cluster basis prior to adaptation. To get a fairer comparison, the gender detection should be done prior to clustering, and clustering should be done within gender. This probably explains why the GD-NW system is only 0.3 points better than the GI-NW system after adaptation. In an early development experiment on the 1996 dev set that used the true turn marks to determine the segment boundaries and used the speaker identities to determine the clusters, we saw a 1 point improvement when moving from a GI-NW system to a GD-NW system, and an additional 1 point improvement when moving from the GD-NW system to a GD-W system. This result was true both before and after speaker adaptation. The GI-W system would probably benefit from using the warp scales in the clustering process, which may explain why we don't get as big an improvement over the GI-NW as we might expect.

	GI-NW		GD-NW		GI-W	
	before	*after*	*before*	*after*	*before*	*after*
F0	18.7	17.4	18.2	17.4	17.6	16.9
F1	31.9	29.6	31.1	30.0	29.7	28.2
F2	48.1	41.4	46.2	40.1	43.8	37.8
F3	41.7	39.5	41.1	39.2	38.9	37.9
F4	37.2	33.3	35.6	33.0	34.2	33.0
F5	35.1	31.6	35.6	32.2	35.3	32.2
FX	55.8	54.4	53.2	49.1	54.6	50.3
Total	30.9	28.1	30.0	27.8	28.9	26.9

Table 4: Comparison of gender-independent non-warped, gender-dependent non-warped, and gender-independent warped acoustic models, both before and after adaptation.

3.3 Automatic Segmentation and Clustering

We next examine how well our automatic segmentation/clustering system performed. In the following experiments we ran a modified version of our evaluation system on the 1996 HUB4 evaluation data. The differences are that the acoustic models are slightly smaller non-SAT models trained only from the 1997 HUB4 acoustic training corpus, using frequency warping, while the language model is a small bigram language model trained only from the transcripts of the 1997 HUB4 acoustic training corpus. To explore the efficacy of our algorithms, we tried replacing the automatically generated segments by the known turn marks, and the automatically generated clusters by clusters determined by the known speaker identities. In Table 5, the 'known/known' sytem uses the known turn marks and speaker identities, the 'known/auto' system uses the turn marks but automatically clusters them, while the 'auto/auto' sytem uses automatic segmentation and clustering.

chop:	known		known		auto	
clust:	known		auto		auto	
	before	*after*	*before*	*after*	*before*	*after*
F0	32.6	30.7	32.6	30.9	32.6	31.2
F1	35.1	34.4	35.6	35.0	37.1	36.2
F2	42.7	37.9	45.3	44.7	48.2	43.2
F3	35.9	32.7	37.1	33.9	39.6	39.0
F4	42.6	39.9	43.3	40.1	45.3	43.0
F5	46.2	45.2	49.5	49.2	44.5	43.8
FX	59.6	56.6	59.5	57.1	61.1	59.5
Tot	38.7	36.7	39.3	37.7	40.5	38.9

Table 5: Performance of automatic segmentation and clustering compared to using known turn-marks and speaker identities.

Let's compare the first two systems, which will give us an estimate of how well the clustering worked with warping and adaptation. We lose 0.6% by automatically clustering prior to rapid adaptation, which grows to a 1% loss after rapid adaptation. Since we warp within clusters, the first loss may be attributed to warping errors due to the clustering. The additional 0.4% loss represents the additional loss from using the clusters instead of the true speaker identities when adapting.

Now let's compare the last two columns, in order to estimate how much we lose from automatic segmentation. Notice that before and after adaptation, we lose 1.2% which is attributable to segmentation errors.

Overall we lost 2.2%: 1% due to warping/clustering errors and 1.2% was due to segmentation errors. By taking into account the warp scales during the clustering we can probably reduce warping/clustering errors.

3.4 Acoustic Modelling

In the following experiments all of the acoustic models were trained from warped training data, the test set was the 1996 dev test (where the speaker side information and turn marks are used for warping and adaptation), and the language model was a small bigram language model trained from the 1997 HUB4 acoustic training corpus transcripts.

Our best system trained from the first half of the HUB4 acoustic training corpus, i.e. the 1996 HUB4 acoustic training corpus, had a word error rate of 40.8%. (We did our initial development using this data in preparation for the STREC evaluation, where we were required to run a recognizer, trained from the first half of the training corpus, on the second half of the training corpus.) When we added the rest of the HUB4 training data, bringing the total amount of training data up to about 70 hours, we saw a 1% absolute reduction in the error rate. When we added word boundary information to our decision tree triphone clustering, we saw a further 1% improvement. Two passes of Baum-Welch adaptation gave an additional small improvement of 0.2%, which we believe is significant since the improvement occurred in each category (F0-FX). These models were used to seed the Speaker Adaptive Training (SAT) process. Results are summarized in Table 6.

	WER
Trained on initial 35 hours	40.8
Trained on entire 70 hour set	39.7
+ use word boundary	38.8
+ 2 passes of BW adapt'n	38.6

Table 6: Word error rate with increased training data and addition of word-boundary information and Baum-Welch adaptation.

We had disappointing results with SAT, which we are still puzzled by. We saw no improvement when just training with HUB4 data. Only after adding data from other corpora did we see any gains, and even then the gains were small. In addition to adding the 1995 Marketplace development data we tried adding material from the Wall Street Journal, Macrophone, and WSJCAM0 corpora. The hope was that by adding WSJ data we would improve the "read studio" category (F0), that by adding Macrophone data we would improve the "low bandwidth" category (F2), and that by adding WSJCAM0 data we would improve the "non-native speaker" (F5) category. In the case of the WSJ and Macrophone corpora, we selected gender balanced subsets for training, consisting of 20 and 10 hours resp., while we used all of the WSJCAM0 data. Since the Macrophone data was sampled at 8 kHz, we upsampled it to 16 kHz and then applied a low-pass filter, before proceeding with our usual front end.

Table 7 summarizes the experiments that we ran, where size refers to the total number of gaussian components in the models, and the "adp WER" column reports the results of unsupervised rapid adaptation using 4 transformations.

The addition of the WSJ data had the biggest impact, since we had already determined in an earlier non-SAT experiment that there was only a tiny improvement from adding Marketplace data. The initially encouraging improvement from adding the WSJCAM0 data prior to rapid adaptation was nearly erased after rapid adaptation. It will come as no surprise that most of the improvement from adding the WSJCAM0 data was realized in the "non-native speaker" category, F5. Unfortunately,

Available training data:

A = HUB4	70 hrs
B = WSJ si250	20 hrs
C = Marketplace	6 hrs
D = Macrophone	10 hrs
E = WSJCAM0	16 hrs

name	trained on	size	WER	adp WER
b8_a	A (no SAT)	134	38.6	36.9
bse_0	A	134	38.7	36.7
bse_1	A B C	162	38.5	36.1
bse_2	A B C D	165	38.4	36.3
bse_3	A B C E	174	37.6	36.0
bse_4	A B C D E	177	38.0	35.9

Table 7: Performance of SAT models using different training data sources.

adding the Macrophone data made things worse overall, and did not give any improvement in the "low bandwidth" category.

We chose to use the bse_3 models for the evaluation system, after subjecting them to an additional two passes of Baum-Welch adaptation. We also tried increasing the allowed number of components per PEL from 32 up to 48, but that did not yield any improvement.

	size	WER	adp WER
bse_3 models	174	37.6	36.0
plus 2 passes of BW	174	37.6	35.6
up to 48 comps/PEL	203	37.6	36.1

Table 8: Further model tuning -- adding 2 passes of Baum-Welch adaptation during training and increasing the number of components per mixture.

3.5 Retuning

During the follow-up analysis of our Mandarin Broadcast News system [12], we discovered that the system was significantly mistuned and that results could be improved by widening recognition thresholds. We tried using these new Mandarin-tuned recognizer settings on the English system and were delighted to see a 1.9% reduction in word error rate over our official evaluation run, as reported in Table 9.

4. FUTURE WORK

Much of our time in preparing for this evaluation was spent simply organizing and learning to work with the vast body of training and test materials that make up the HUB4 corpus. Now that we have a Broadcast News transcription system in place at last, we look forward to embarking on a number of experiments. Among other projects, we will be studying more

	official	tuned
F0	14.9	12.9
F1	23.4	22.2
F2	31.9	30.2
F3	35.8	33.4
F4	30.0	27.7
F5	21.3	18.4
FX	45.7	43.0
Tot	23.3	21.4

Table 9: Re-running the evaluation system with new recognizer settings.

carefully the interaction of the clustering with speaker normalization. We will also explore adding easily obtained and reliable side information, such as gender, to improve the clustering process.

Acknowledgements

A number of individuals assisted in the preparation of materials for this evaluation system. We would particularly like to thank Eric Wheeler for his work during summer 1997 in reviewing and organizing the training and test material and for his help with the early clustering experiments, and Kristin Baldwin for her help in creating pronunciations for the lexicon.

This work was supported by the Defense Advanced Research Projects Agency. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the U.S. Government and no official endorsement should be inferred.

REFERENCES

[1] S. Wegmann et al., "Marketplace Recognition using Dragon's Continuous Speech Recognition System," *Proc. DARPA Speech Recognition Workshop*, Arden House, February 1996.

[2] B. Peskin et al., "Progress in Recognizing Conversational Telephone Speech," *Proc. ICASSP-97*, Munich, April 1997.

[3] R. Roth et al., "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, 1995.

[4] H. Hermansky, "Perceptual Linear Prediction (PLP) Analysis for Speech," *J. Acoust. Soc. Amer.*, vol. 87, 1990, pp. 1738-1752.

[5] P. Woodland et al., "Broadcast News Transcription Using HTK," *Proc. DARPA Speech Recognition Workshop*, Chantilly, February 1997.

[6] M.J. Hunt et al., "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination," *Proc. ICASSP-91*, Toronto, May 1991.

[7] S. Wegmann et al., "Speaker Normalization on Conversational Telephone Speech," *Proc. ICASSP-96*, Atlanta, May 1996.

[8] C.J. Leggetter and P.C. Woodland, "Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression," *Proc. ICSLP'94*, Yokohama, September 1994.

[9] V. Nagesha and L. Gillick, "Studies in Transformation Based Adaptation," *Proc. ICASSP-97*, Munich, April 1997.

[10] T. Anastasakos et al., "A Compact Model for Speaker-Adaptive Training," *Proc. ICSLP'96*, Philadelphia, October 1996.

[11] F. Kubala et al., "The 1996 BBN Byblos Hub-4 Transcription System," *Proc. DARPA Speech Recognition Workshop*, Chantilly, February 1997.

[12] P. Zhan et al., "Dragon Systems' 1997 Mandarin Broadcast News System," *these Proceedings*.